LOCAL QUALITY ASSESSMENT FOR OPTICAL COHERENCE TOMOGRAPHY

Peter Barnum *Mei Chen* *Hiroshi Ishikawa* *Gadi Wollstein* *Joel Schuman*

Robotics Institute
Carnegie Mellon University

Intel Research Pittsburgh

UPMC Eye Center, Department of Ophthalmology
University of Pittsburgh School of Medicine

ABSTRACT

Optical Coherence Tomography (OCT) is a non-invasive tool for visualizing the retina. It is increasingly used to diagnose eye diseases such as glaucoma and diabetic maculopathy. However, diagnosis is only possible when the layers of the retina can be easily distinguished, which is when the images are evenly illuminated. Automated OCT quality assessment (i.e. signal strength) is only available for images as a whole. In this work, we present an automated method for *local* quality assessment. For training data, three OCT experts label the quality of each individual a-scan line in 270 OCT images. We extract features that are insensitive to pathology, and employ a hierarchy of support vector machines and histogram-based metrics. Our trained classifier is able to determine not only when signal strength is low, but also when it will affect doctors' diagnostic ability. Our results improve over the state of the art in OCT quality assessment.

Index Terms— Image quality assessment, optical coherence tomography

1. INTRODUCTION

Optical Coherence Tomography (OCT) is a powerful tool for imaging the retina in vivo [1]. It uses the properties of coherent light interference to image at an axial resolution of about 8 microns. This allows for diagnosis and assessment of diseases such as glaucoma and diabetic maculopathy. Since its introduction in 1991, OCT has become increasingly popular in hospitals around the world.

If an OCT image has low signal strength, then it is difficult to see the eye's physiology, making correct diagnosis difficult. Quality for whole images can be determined automatically [2], but as seen in Fig. 1, sometimes only a portion of the image is bad. In current clinical practice, an image is discarded if even a small part is difficult to see. This means that more images need to be taken, which is time consuming for the doctor and troublesome for the patient. But if it is known which sections are high or low quality, then only the completely useless images would need to be discarded. It might even be possible to create a composite from the good parts of several images.

An OCT image is a collection of one dimensional depth samples (a-scans). The reflectivity of the tissue at each depth along the sample line is recorded. To facilitate interpretation, a false color scheme is used for all images in this paper. From highest to lowest tissue reflectivity, the colors are white, red, yellow, green, blue, then black.

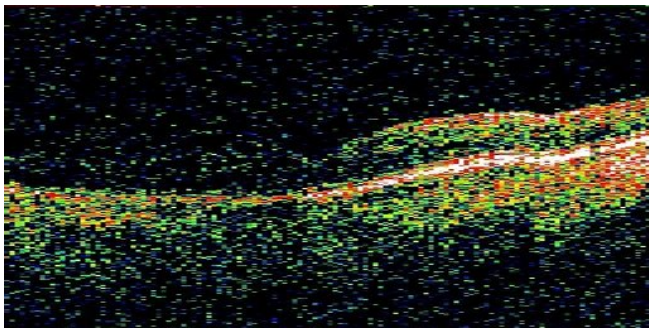


Fig. 1. The quality of OCT images can vary within a single image. An image does not necessarily need to be discarded, if only part of it is illegible. In this image, even though the left part is low quality, the right part is excellent and all retinal layers can be seen.

We propose a hierarchical support vector machine (SVM) based method for computing the quality of individual a-scans. The SVM is trained on data labeled by three experts. This automated quality estimation could potentially be used to guide an image compositing or segmentation algorithm. Our results show that this method outperforms the state of the art in OCT quality assessment.

2. BACKGROUND AND RELATED WORK

In this paper, we are primarily concerned with quality in terms of image intelligibility rather than fidelity [3]. In other words, factors such as the brightness or level of noise are unimportant, unless they affect diagnostic accuracy.

Various factors affect OCT image quality. Somfai et al. [4] discuss common causes of poor quality OCT images: defocus, depolarization, and improper centering. There can also be more subtle problems with incorrect retinal thickness measurements [5, 6, 7]. Since this type of poor quality cannot be determined from a single image, it is not a component of our automated quality assessment.

OCT machines assess quality of images as a whole, reporting overall signal to noise ratio and signal strength. Stein

The first author performed this work while at Intel Research Pittsburgh

et al. [2] developed a more clinically accurate global quality assessment algorithm. In this paper, we will build off the whole-image quality assessment of Stein et al., and determine the quality of individual image regions.

3. EXPERT DATA LABELING

The goal of this paper is to determine image quality independent of pathology. Therefore, instead of considering only healthy subjects, we selected a mix of healthy and diseased eyes. Thirty each have no glaucoma, early glaucoma, and advanced glaucoma. The level of glaucoma was determined with a Humphrey visual field glaucoma hemifield test, intraocular pressure, and the appearance of the optic nerve head. The threshold to distinguish between early and advanced glaucoma was selected to be a mean deviation of -9 dB on the Humphrey visual field.

For each subject, we used one image each of the macula, optic nerve head (ONH), and a peripapillary circular scan imaging the retinal nerve fiber layer (NFL). Three OCT experts each labeled the quality of every a-scan in all 90x3 images. As in [6, 7], we defined three levels of quality, *excellent*, *acceptable*, and *poor*. For this study, quality refers to the signal strength relative to the best possible, ignoring intrinsic limitations of OCT. We wanted to determine the usefulness of the image, independent of unavoidable artifacts. Four specific examples of unavoidable artifacts (shown in Fig. 2) are shadowing, anything causing a wave or discontinuity in the image (such as eye movement), pathology, and individual differences. The experts would only label an image as poor if there was low signal strength independent of these effects.

To determine intra-operator variability, each expert labeled 30 of the images twice. Ground truth is defined as the mode of the three if it exists, otherwise it is the median across experts. The difference between *acceptable* and *excellent* is subtle. Therefore, to train and evaluate our algorithm, we used the label *good* for both, reducing the problem to differentiating between *good* and *poor* a-scans. The experts' quality assessment is discussed in the results in Section 5.

4. ALGORITHM

We aim to determine the quality for each individual a-scan. But often it is difficult to determine the quality of one without looking at its neighbors. For example, a blood vessel can create a shadow that make a small region appear to be of poor quality, although the region looks fine in a larger context. To prevent confusion due to such local effects, while still allowing for per-line classification, a multi-scale analysis is used. Features are extracted from various sized neighborhoods centered around a specific a-scan. The quality of each level of the hierarchy is computed independently, then the estimates are combined to yield a score that is both local and robust to many types of variation.

4.1. Selecting Good Features

We begin by extracting features that are not affected by common pathologies or eye movement. Pathology, such as epiretinal membrane, macular holes, or cystoid macular edema, cause variations that are independent of the skill of the operator and the capabilities of the machine. Two of the most common changes caused by pathology are thinning and thickening of local areas. Thinning occurs when there are a large number of cell deaths, as in glaucoma. In diabetic maculopathy, fluid accumulates in the retinal tissue causing thicker appearance. In addition, no matter how the images are taken, cupping in the ONH and blood vessels create shadows, which results in low reflectivity in local regions. Also, if patients move their eyes during acquisition, then the resulting images may appear discontinuous. And there is natural variation in the retinas structure between individuals, especially in the ONH, but this is considered to be independent of the images' quality.

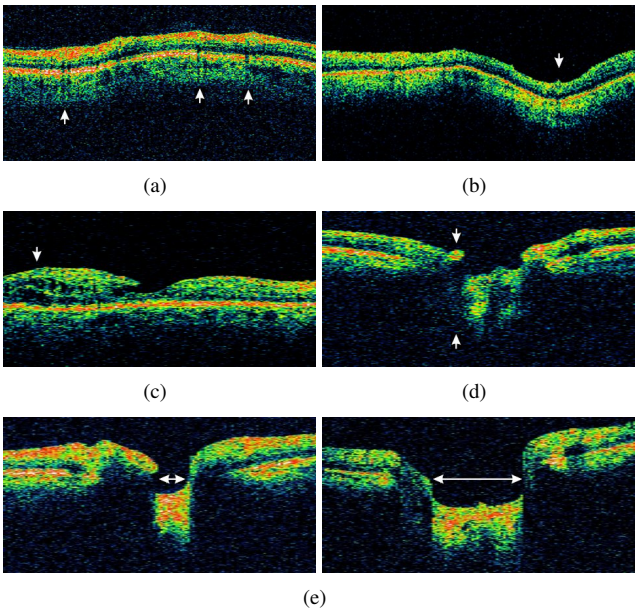
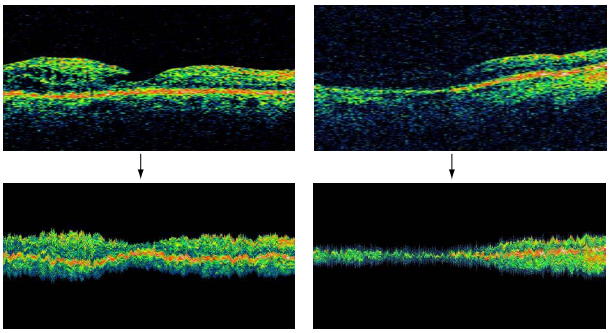


Fig. 2. Since they are unavoidable, we ignore (a) dark areas due to vessel shadowing, (b) waves in the image, (c) retinal thickening, (d) any other eye pathology or shadowing, and (e) individual differences.

It would be possible to employ machine learning to find features that are invariant to these effects, but as in many medical imaging problems, data is scarce. A close examination of the factors in Fig. 2 reveals that most of the variations are types of translation. For example, in Fig. 2 (c), the thickening is simply the separation of retinal layers. Therefore, we use features that are robust to local translation, but still encode much of the spatial structure. As is discussed in more detail in Section 4.2, we independently consider neighborhoods of between 1 and 256 a-scans, each with 1024 depth samples, centered in a specific area, (i.e. we consider one scan, then the one scan and its two neighbors on each side, then one scan and its eight neighbors, etc). In order to run with reasonable memory usage and execution time, we use the Quality



(a) (b)

Fig. 3. Two examples of compression and centering. The edema in (a) is removed without otherwise affecting the image, while the low quality section in (b) is preserved.

Index (QI) score, which is known to linearly correlate with the commercially available signal strength measure [2] for neighborhoods of over five scans. Although not as accurate as the SVM prediction on small neighborhood sizes, the QI gives a good estimate with little computing time or memory usage.

Neighborhoods of under five a-scans are normalized. To begin, we remove noise by setting all samples below percentile p to zero. As is commonly done for OCT images, $p = 75\%$. Next, we compress all non-zero samples together, i.e. we move the first to the top of the image, the second to the spot second from the top, etc. Lastly, the compressed samples are moved so that the mean location of the samples is in the center of the image. This normalization removes variation due to eye movement and retinal thickening. An example is shown in Fig. 3.

4.2. Learning Quality

Each of the three scan types (macula, ONH, and NFL) is trained and tested separately, with leave-one-image-out cross validation (i.e. for 90 images, there are 90 trials). For each of the three types, the quality of each neighborhood size is predicted independently, then combined to determine the final score. When training, if the labeling of a given neighborhood is inconsistent between experts, the most common value is used. For testing, prediction accuracy is defined per a-scan, so no additional processing is required to calculate accuracy.

For the 128x1024 Macula and ONH scans, neighborhoods of $[1, 5, 17, 65, 128]$ a-scans were used. For the 256x1024 NFL scans, $[1, 5, 17, 65, 256]$ were used.

A SVM is trained separately on neighborhood sizes 1 and 5, using the features extracted in Section 4.1, with a radial basis function kernel. For each of the two SVMs, the probability is calculated by fitting a sigmoid to a 3-fold cross-validation of the training set [8]. For the QI scores, no probability is estimated, therefore $P(ascan = good|b_n) \in \{0, 1\}$, where b_n is a neighborhood of n scans.

Given the small amount of data, it would be difficult to determine the full joint probability of all neighborhood sizes. Instead, an independence assumption is made, giving

$$P(ascan = good|b_1, b_5, ...) = \prod_i P(ascan = good|b_i) \quad (1)$$

The probability is then used as a threshold to find the sensitivity at different specificities.

Wollstein's Labeling

	poor	*acceptable*	*excellent*	
poor	12.7	3.2	0.0	3-class repeatability: 94.57%
acceptable	2.2	81.9	0.0	2-class repeatability: 94.57%
excellent	0.0	0.0	0.0	

Ishikawa's Labeling

	poor	*acceptable*	*excellent*	
poor	11.3	1.7	0.0	3-class repeatability: 87.00%
acceptable	4.8	21.4	12.2	2-class repeatability: 96.05%
excellent	0.0	1.9	46.6	

Schuman's Labeling

	poor	*acceptable*	*excellent*	
poor	5.4	3.8	0.0	3-class repeatability: 79.37%
acceptable	0.1	44.3	8.6	2-class repeatability: 93.52%
excellent	0.0	0.4	37.2	

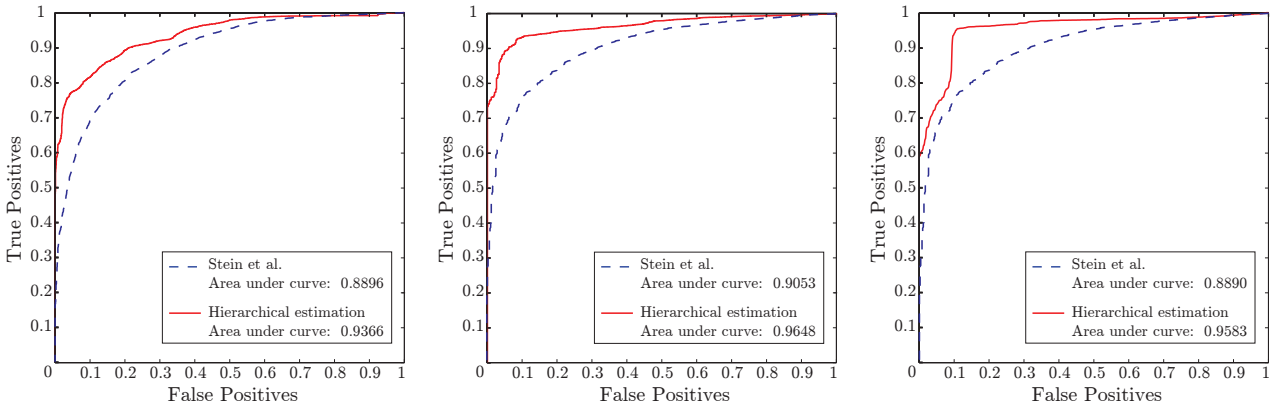
Fig. 4. Analysis of intra-operator variability. Each of the three OCT experts labeled thirty images twice. The charts show the difference between the two labellings. (For example, Wollstein labeled 3.2% of the a-scans as acceptable in the first trial and poor in the second). Repeatability is the percentage of a-scans that were given the same quality label both times, for both three classes (excellent, acceptable, or poor) and two classes (good or poor).

5. EXPERIMENTAL RESULTS

In this section, we examine the experts' labeling in more detail and evaluate the accuracy of our algorithm. To determine intra-operator variability, a set of thirty images was selected, with ten images each of the macula, NFL, and ONH. The set was selected to have approximately equal numbers of excellent, acceptable, and poor quality images. Fig. 4 displays the percentage of each quality class. If they were completely consistent, then the diagonal would sum to 100%.

To determine inter-operator variability, we calculate each expert's accuracy at predicting the others' labellings, shown in Fig. 6. In this case, the two classes are *good* and *poor*. For example, if one expert labeled an image as entirely poor, but another labeled only half as good, then there would be 50% agreement between them. We also include the mode estimate and the results from our algorithm.

Fig. 5 shows ROC curves comparing our work to [2]. For our algorithm, after a certain point, it takes a great deal of false positives to increase the true positive rate. This is likely due to inconsistent quality assignments in the ground truth. Also note that the curves for [2] are fairly smooth. This is likely because the QI does not generalize sufficiently.



(a) Macula (b) Nerve Fiber Layer (NFL) (c) Optic Nerve Head (ONH)

Fig. 5. ROC curves and Area Under Curve (AOC) for prediction accuracy of our algorithm compared with Stein et al. [2].

	Wollstein	*Ishikawa*	*Schuman*	*Mode*	*Algorithm*
Wollstein	–	93	94	97	93
Ishikawa	93	–	92	95	95
Schuman	94	92	–	97	92
Mode	97	95	97	–	95
Algorithm	93	95	92	95	–

Fig. 6. Confusion matrix for inter-operator variability, for each of the three experts, their mode, and the algorithm presented in this paper. Shown is the percentage of scans labeled the same, (e.g. Schuman was 94% consistent with Wollstein). In all cases, the algorithm was trained on the mode.

6. CONCLUSION

We have presented an automatic algorithm that estimates the local quality of OCT images, in a way that is insensitive to pathology. We first train SVMs and use the QI metric independently for different sized neighborhoods of a-scans, then combine the individual estimates. This hierarchical method is significantly more accurate than the state of the art in OCT quality estimation. For future work, this method can be extended to explicitly model pathology and individual differences, and to work with volumetric measurements from a spectral OCT. Accurate quality assessment will decrease the time patients have to spend being imaged, reduce doctors workload, and improve the accuracy of medical image processing algorithms.

7. REFERENCES

[1] D. Huang, E. Swanson, C. Lin, J. Schuman, W. Stinson, W. Chang, M. Hee, T. Flotte, K. Gregory, C. Puliafito, and J. Fujimoto, "Optical coherence tomography," *Science*, vol. 254, pp. 1178–81, 1991.

[2] D.M. Stein, H. Ishikawa, R. Hariprasad, G. Wollstein, R.J. Noecker, J.G. Fujimoto, and J.S. Schuman, "A new quality assessment parameter for optical coherence tomography," *British Journal of Ophthalmology*, vol. 90, pp. 186–190, 2006.

[3] W.K. Pratt, *Digital Image Processing: PIKS Inside*, Wiley-Interscience, 3rd edition, 2001.

[4] G.M. Somfai, H.M. Salinas, C.A. Puliafito, and D.C. Fernández, "Evaluation of potential image acquisition pitfalls during optical coherence tomography and their influence on retinal image segmentation," *Journal of Biomedical Optics*, vol. 12, no. 4, 2007.

[5] M. Sehi, D.C. Guaqueta, W.J. Feuer, and D.S. Greenfield, "A comparison of structural measurements using 2 Stratus optical coherence tomography instruments.," *Journal of Glaucoma*, vol. 16, no. 3, pp. 287–92, 2007.

[6] M.E.J. van Velthoven, M.H. van der Linden, M.D. de Smet, D.J Faber, and F.D Verbraak, "Influence of cataract on optical coherence tomography image quality and retinal thickness," *British Journal of Ophthalmology*, vol. 90, pp. 1259–1262, 2006.

[7] D. M. Stein, G. Wollstein, H. Ishikawa, E. Hertzmark, R.J. Noecker, and J. S. Schuman, "Effect of corneal drying on optical coherence tomography," *Ophthalmology*, vol. 113, no. 6, pp. 98591, 2006.

[8] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds., pp. 61–74. MIT Press, 1999.